Poor access to Apanteles[1] species literature through titles, abstracts and automatically extracted species anmes as keywords[2]

Lee J. Shervis and R. D. Shenefelt[3]

Department of Entomology, University of Wisconsin, Madison 53706

ABSTRACT

A collection of 183 documents on 3 species of *Apanteles* was examined for 1) frequency of the species names in titles, 2) frequency of the species names in published abstracts, and 3) variations in spelling of the species names in the original texts. These species were mentioned in 0-3% of the titles and in 0-29% of 97 published abstracts, suggesting the need for greater depth of analysis of the literature. Numerous variations in spelling of the genus, species and author components of the species names were encountered in the full texts, creating a special problem in the use of wholly automatic full text processing and searching.

Titles and abstracts, no matter how carefully written, cannot convey the entire content of a long or complex document. An information retrieval system which indexes only titles or abstracts, therefore, omits access to information which does not comprise a large percentage of the full document. Automatic full text processing techniques, such as those discussed by Wilson (1966) have been suggested as a solution to this problem. This technology involves the mechanical conversion of the printed or microfilmed page to machine readable form, automatic extraction of keywords from full documents and automatic retrieval of information from the resulting files through keywords selected by the user.

A cursory examination of Braconidae literature showed that information on a given species, in spite of its absolute length, frequently comprised too small a proportion of a complete text to be treated in an abstract. It was also noticed that the scientific names of the species were frequently misspelled in the literature, presenting an apparent obstacle to efficient retrieval through automatic full text processing and searching. The purpose of this study was to obtain quantitative data on the above problems, using a selected portion of Braconidae literature as the working base. The study was part of preliminary work toward the construction of a prototype Braconidae information retrieval system.

Methods

A collection of 183 documents[4] on 3 species of *Apanteles* was examined for 1) frequency of the species names in titles, 2) frequency of the species names in published abstracts, and 3) variations in spelling of the species names in the original texts. The species studied were *Apanteles melanoscelus* (Ratzeburg) 1844, *A. porthetriae* Muesebeck 1928, and *A. ocneriae* Ivanov 1899, all parasites of the gypsy moth, *Porthetria dispar* (L.). 167 of the documents dealt with *melanoscelus,* 38 with *porthetriae* and 14 with *ocneriae*. Twenty-seven of the documents contained information on 2 or all of these species.

[1]Hymenoptera: Braconidae

[2]Research supported by the College of Agricultural and Life Sciences, University of Wisconsin, Madison, and by means of a cooperative agreement between the College and the Agricultural Research Service, USDA.

[3]Specialist and professor, respectively.

[4]The bibliography will be published in the June, 1973 issue of this *Journal.*

The documents consisted of journal articles, U.S. federal, U.S. state and foreign government publications, conference proceedings and technical books (catalogs, manuals, textbooks, etc.) in 12 languages. Species information from a document ranged in length from a single sentence or footnote to over 30 pages. The documents themselves were from 1-2 pages to 1400 pages in length.

The Review of Applied Entomology (Series A: Agricultural) was used to obtain abstracts of these documents. It was selected as the best single source for our subject matter because of its extensive and early (1913-present) coverage of the world literature. Abstracts for 97 of our documents were obtained, located through the author index. The documents pertaining to *melanoscelus, porthetriae* and *ocneriae* were represented by 93, 16 and 3 abstracts, respectively. The number of abstracts obtained did not reflect the relative coverage of *RAE;* some of the 183 documents were published before 1913, others were men-

Table 1.—Variations in spelling of scientific names—*A. melanoscelus, A. porthetriae, A. ocneriae.*

I. *Apanteles melanoscelus* (Ratzeburg) 1844 (167 documents)			
Genus variations:	A	*Author variations:*	(blank)
	Ap		Latr
	Apantales		Ratz
	Apanteles		Ratzb
			Ratzbg
Species variations:	m		Ratzeberg
	malanoscelus		Ratzeburg
	melanocelis		Rbg
	melanocelus		Ritg
	melanocephalus		Rtz
	melanoscelis		Rtzb
	melanoscellus		Rtzbg
	melanoscelus		
	melanoschelus		
	melanoseclis		
	melinosus		
II. *Apanteles porthetriae* Muesebeck 1928 (38 documents)			
Genus variations:	A	*Author variations:*	(blank)
	Apanteles		L
			Mues
Species variations:	portehtriae		Muesb
	portethriae		Muesebeck
	portheriae		Mus
	porthertiae		Nees
	porthetria		new species
	porthetriae		
	porthretiae		
	porttetriae		
III. *Apanteles ocneriae* Ivanov 1899 (14 documents)			
Genus variations:	A	*Author variations:*	Iv
	Ap		Ivan
	Apanteles		Ivanov
			Ivanow
Species variations:	ochneriae		Iw
	ocneria		Iwanov
	ocneriae		Iv
			Stanov
			Svan
			Svanow
			Tw

tioned by title only, a few were missed because of publishing time lag, and one year's volume was not available.

Results

Titles. — Only 4 of the 167 documents containing information on *A. melanoscelus* (3%) mentioned this species in the title. None of the *A. porthetriae* or *A. ocneriae* documents contained the species names in the titles.

Abstracts.—The species was mentioned in 27 of the 93 abstracts of *A. melanoscelus* documents, in 2 of 16 for *A. porthetriae* and in none of 3 for *A. ocneriae*, or 29%, 12.5% and 0%, respectively. The percentage probably would not have improved had a complete set of abstracts been available, since most of the documents mentioned only by title were long works in which the species information was relatively brief.

The low frequency of the species names in titles and abstracts suggested the need for greater depth of analysis of the literature to achieve thorough access to species-level information.

Spelling of Species Names.—The misspelling of words in original documents is seldom mentioned as an important problem in automatic full text processing, and it probably isn't for normal words. Our sample collection, however, revealed a rather significant variety of spellings of the names of the 3 *Apanteles* species. These variant spellings apparently were due to 1) alternate interpretations of the Latin grammar rules in the International Code of Zoological Nomenclature, 2) inconsistent abbreviations of the genus and/or author components, 3) the failure of authors to adequately verify spellings prior to publication, and 4) typographical errors.

The spellings encountered are listed in Table 1. Synonyms were not recorded, nor were variations in punctuation and capitalization, e.g. "?Apanteles melanoscelus", "L'Apanteles melanoscelus", "A. Por-thetriae", "(Ratz)".

The accepted spellings "melanoscelus", "porthetriae" and "ocneriae" failed to occur at least once in 13%, 19%, and 21% of the respective documents.

The combination of the correctly spelled species term plus the most common stem for the author component ("Ratz--", "Mues--", "Ivan--") failed to occur in 36%, 24% and 64% of the respective documents.

Conclusions and Discussion

The occurrence of spelling variations in technical entomological literature creates a special problem worth noting, in spite of the fact that automatic full text processing is perhaps not a viable alternative for efficiently handling the existing literature. The recent "Wigington Report" (1972) states that technical problems still exist in converting the printed page to machine readable form without keyboarding its content. One such problem is the high error rate of present optical character recognition equipment when multiple font recognition is required.[5] A more fundamental problem, however, is the well known one of inherently poor recall-precision levels.

It isn't our intention to dismiss any possible alternative that may ultimately aid in coping with the information problem, or parts of it, but we thought it desirable to point out some of the problems of particular importance in handling the existing literature of the 1½ million described species of insects.

A manual literature processing technique has been developed (Shervis et al., 1972) which offers the potential for virtually complete recall of published species information with acceptable levels of noise. It is hoped that this technique will eventually prove useful in handling portions of the existing body of entomological literature.

Acknowledgment

Our appreciation is expressed to Dr. Richard H. Foote for reading the manu-

[5]It was tempting to illustrate the remarkable variety of fonts encountered in our collection; in one document (Crossman, 1922) the species name alone occurred in 12 different fonts.

script and for many suggestions regarding the research.

References Cited

Crossman, S. S. 1922. *Apanteles melanoscelus*, an imported parasite of the gipsy moth. U.S. Dep. Agr. [Dep.] Bull. 1028: 1-25.

Shervis, L. J., R. D. Shenefelt, and R. H. Foote. 1972. Species-level analysis of biological literature for storage and retrieval. BioScience 22: 651-655.

[Wigington Report] National Academy of Sciences. Computer Science and Engineering Board. Information Systems Panel. 1972. Libraries and information technology; a national system challenge. Washington, D.C., Nat. Acad. Sci. xi + 84p.

Wilson, R. A. 1966. Optical page reading devices. New York, Reinhold. ix + 197p.

NOTICE

1973 Programs

John Wesley Powell Auditorium
Cosmos Club, 2170 Florida Ave., N.W., Washington, D.C.

8:00 P.M. Public Welcome

April 19
Dr. Max V. Matthews Acoustical and Behavioral Research Center, Bell Telephone Laboratories

Computer Music and Other Unusual Computer Applications

This subject will intrigue those who are interested in using computers for functions other than performing arithmetic operations very rapidly or handling masses of data. The speaker will demonstrate how useful a computer could be to a composer, for instance, if he had a program available which took his score and produced a tape on which the composition has been "performed" by the computer and which he can play on his tape recorder. In addition, he will discuss one or more of the other unusual uses to which he has been putting computers, such as an aid to composing music, and for typesetting.

May 17
Dr. Richard K. Cook President, Washington Academy of Sciences

Annual Dinner Meeting

For information contact:

Washington Academy of Sciences Office
9650 Rockville Pike (Bethesda), Washington, D.C. 20014
Telephone:530-1402

PARKING: Available at the Cosmos Club, on the street, and at the Fairfax Hotel across from the Cosmos Club (2121 Massachusetts Ave.)